

科学技術日本語および英語の語彙調査

小 森 早 江 子*
木 戸 哲 夫**

A Study of Lexical Analysis for Teaching
Technical Japanese and English

Saeko KOMORI*

Tetsuo KIDO**

1 はじめに

近年、大学レベルにおける語学教育で目的別語学教育 (Japanese for Specific Purposes や English for Specific Purposes) が注目をあつめるようになっている。科学技術日本語および科学技術英語の授業の目的は、学生が自分の専門分野の学習を進める上で十分な日本語力や英語力を身につけることである。専門分野で使われる日本語や英語を教育するためには、専門分野で使われる日本語や英語が一般的な日本語や英語と比べて、語彙の面でどのように異なるのか知ることがそれぞれの違いに応じた効果的な教授法を探る手がかりになる。このような考えに基づいて目的別語学教育のための効果的な教授法への第一歩として、本研究では科学技術日本語および科学技術英語の特徴を計量的に明らかにしようとした。分析の対象としたデータは、日本語は朝日新聞の社説と工業新聞の記事、英語は The Lancaster/Oslo-Bergen Corpus (以下、LOB コーパスと略す) を使用して、一般の文章と科学技術系の文章を比較した。なお、日本語には、英語のように研究者が利用できる言語分析のために用意された各種のコーパスというものがないため、今回分析の対象とした資料も市販の CD-ROM とパソコン通信で得た新聞記事を使用した。

2 科学技術日本語の分析

日本語のデータとしては、すでに電子化された朝日新聞の社説および工業新聞の記事それぞれ10日分をデータとして使用した¹⁾。朝日新聞の社説のデータは約64KB、工業新聞の記事のデータは約16KBである²⁾。

日本語は、現在漢字仮名まじりの表記法をとっている。科学技術日本語に、その表記において

*外国語教室 (Department of Foreign Languages)

**外国語教室 (Department of Foreign Languages)

際だった特徴があるのか、あるとすればそれはどのような点なのかを調査することにした。さらに文章理解の基礎となる動詞および助動詞に着目して、科学技術日本語の特徴を調べる語彙調査をした。

日本語の分析のためには、対象とする文章の形態素解析が必要である。本研究では、NTT 情報通信処理研究所で開発された形態素解析プログラム INDEXER を利用した。INDEXER は、入力した日本語の文を形態素解析し、一行ごとに形態素とその品詞コードを出力するプログラムである³⁾。例 1 に示すのは、INDEXER の出力結果の一部である。対象としたテキストは、朝日新聞の社説「国際政治と日本の体質強化」(1991年1月1日)の冒頭の部分である⁴⁾。

〈例 1〉⁵⁾

宇宙	(ウチュー)	1100
から	(カラ)	7410
見	(ミ)	2410
る	(ル)	2417
地球	(チキュー)	1940
に	(ニ)	7430
は	(ワ)	7530
、	()	0120
どこ	(ドコ)	1720
に	(ニ)	7430
も	(モ)	7530
線	(セン)	1100
は	(ワ)	7530
引っ張	(ヒッパ)	2380
っ	(ッ)	2384
て	(テ)	7630
い	(イ)	2811
ない	(ナイ)	7186
。	()	0110

2.1 語彙の分析

日本語の文章は、漢字、ひらがな、カタカナ、数字が使われている。当然のことながら、日本語学習者が日本語で書かれた文章を読み、理解するためには、この4つの文字のそれぞれを知る必要がある。

まず、語彙の種類の調査として、上のように解析した朝日新聞の社説と工業新聞の記事の形態素を、それぞれ漢字、ひらがな、カタカナに分類し、それぞれの比率を比較した。その結果を表 1 に示す。ここでは、漢字のみまたは漢字を含む語彙を「漢字語彙」、ひらがなのみあるいはひ

〈表 1〉

	朝日新聞の社説 (%)	工業新聞の記事 (%)
1 漢字語彙	38.0	44.3
2 ひらがな語彙	65.2	49.6
3 カタカナ語彙	2.2	10.4

らがなを含む語彙を「ひらがな語彙」、カタカナのみあるいはカタカナを含む語彙を「カタカナ語彙」と呼ぶことにする。なお、漢字語彙、ひらがな語彙、カタカナ語彙のそれぞれは、重複するものがあるため3つの合計は100%を越える値となる。

表1を見ると、朝日新聞の社説も工業新聞の記事も、比率として一番多いのは、ひらがな語彙で、次に多いのは、漢字語彙である。両者の比較で特徴的なことは、カタカナ語彙が朝日新聞の社説2.2%に対して、工業新聞の記事では10.4%と4.7倍以上使われていることである。

さらに、漢字語彙、ひらがな語彙、カタカナ語彙の内訳を示したものが、表2である。つまり表1の漢字語彙は、漢字のみ、漢字+ひらがな、漢字+カタカナ、カタカナ+漢字、ひらがな+漢字を合計したもの（表2の1，4，5，6，7，9）で、表1のひらがな語彙は、ひらがなのみ、漢字+ひらがな、ひらがな+漢字、カタカナ+ひらがなを合計したもの（表2の2，4，7，8）で、表1のカタカナ語彙は、カタカナのみ、漢字+カタカナ、カタカナ+漢字、カタカナ+ひらがなを合計したもの（表2の3，5，6，8）である。

〈表2〉

表記の種類 (例)	朝日新聞の社説 (%)	工業新聞の記事 (%)
1 漢字のみ (欧州)	32.3	38.9
2 ひらがなのみ (いまに)	60.1	45.7
3 カタカナのみ (イラク)	1.9	9.1
4 漢字+ひらがな (本当に)	4.9	3.7
5 漢字+カタカナ (東南アジア)	0.1	1.2
6 カタカナ+漢字 (ソ連)	0.2	0.0
7 ひらがな+漢字 (つり上げ)	0.2	0.1
8 カタカナ+ひらがな (クギづけ)	0.0	0.1
9 数字+漢字 (10月)	0.3	0.4
10 アルファベット (TBS)	0.2	2.0

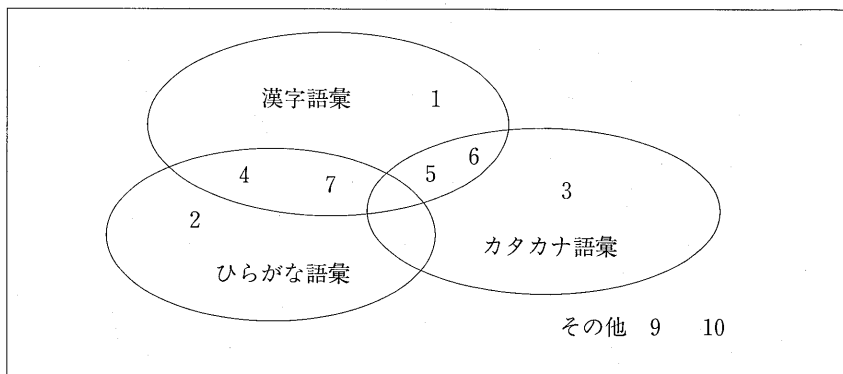


表2でも、カタカナのみの語彙が朝日新聞の社説1.9%に対して、工業新聞の記事では9.1%と非常に多く見られることがはっきりとわかる。

表1および表2から以下の3つのことがわかった。

- 1) 朝日新聞の社説と工業新聞の記事との間に、漢字語彙の使用頻度においてはあまり差は見られない。
- 2) 工業新聞の記事では、ひらがな語彙が比較的少ない。
- 3) 工業新聞の記事には、カタカナ語彙が4.7倍以上も多く使われている。

つまり、朝日新聞の社説と工業新聞の記事の比較で観察された特徴として、科学技術系の文章では、カタカナ語彙が多く使われる傾向があるということが確認できた。

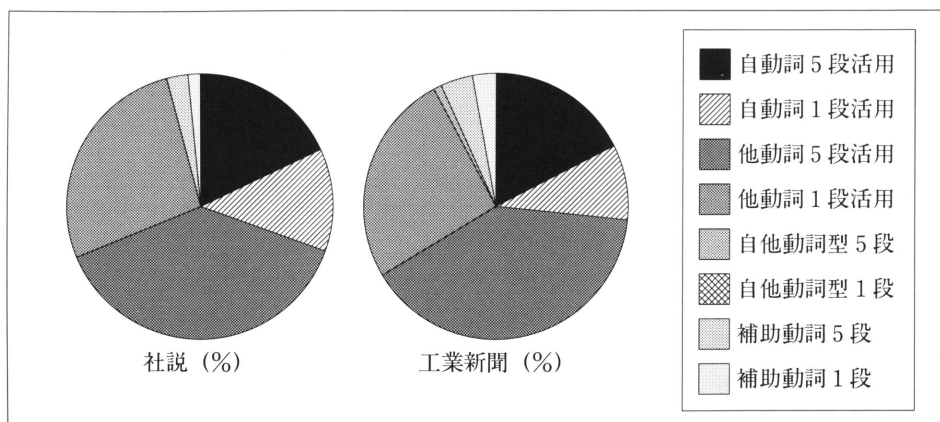
科学技術系の文章には、専門用語が多用され、その中には外来語が多いことから比率としてカタカナ語彙が多く見られるだろうという予測を持っていたが、全体の一割弱という結果は、確かに大きいといえる。外国人学習者にとって外来語は、語源が予測できず、むずかしいものである。効率的な教授法を探るために、さらに語源との関連や学習の困難なものとそうでないものの調査など、詳しく調べる必要がある。

2.2 動詞および助動詞の分析

上の朝日新聞の社説と工業新聞の記事の比較では語彙の面で差が見られた。次に、この2つ

〈表3〉

	朝日新聞の社説 (%)	工業新聞の記事 (%)
1 自動詞 5 段活用	18.2	17.6
2 自動詞 1 段活用	12.4	9.1
3 他動詞 5 段活用	38.3	40.1
4 他動詞 1 段活用	26.9	25.7
5 自他動詞型 5 段	0.0	1.0
6 自他動詞型 1 段	0.0	0.0
7 補助動詞 5 段	2.5	4.3
8 補助動詞 1 段	1.8	2.7



の文章の中核をなす動詞と助動詞に注目して、科学技術日本語に特徴があるかどうか比較する。方法は、INDEXER による形態素解析の結果から、動詞を 8 種類に分類し、それぞれの動詞全体に占める割合を算出した。結果は表 3 に示すとおりである。

表 3 でわかるように、どちらも他動詞が 6 割以上を占めており、朝日新聞の社説と工業新聞の記事のデータにそれほどきわだった差は見られない⁶⁾。

次に助動詞について調査した。これも同様に INDEXER を使って形態素解析をし、受身、使役など、9 種類に分類し、それぞれ助動詞全体に占める割合を算出した。

〈表 4〉

	朝日新聞の社説 (%)	工業新聞の記事 (%)
1 受身 (れる、られる)	18.0	27.6
2 使役 (せる、させる)	8.0	13.8
3 希望 (たい、たがる)	6.0	3.4
4 打消 (ない、ぬ)	20.0	10.3
5 過去 (た、だ)	10.0	10.3
6 断定 (ます、です、だ)	14.0	13.8
7 推量 (う、よう、他)	8.0	6.9
8 伝聞 (そうだ、他)	2.0	0.0
9 様相 (らしい、他)	14.0	13.8

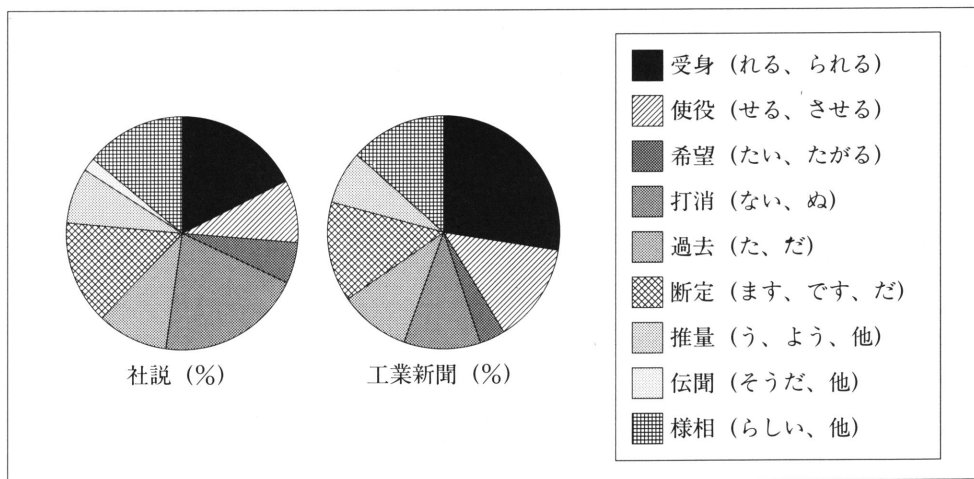


表 4 からわかるように、助動詞については朝日新聞の社説と工業新聞の記事との間にかなり違いが見られた。まず、受身（「れる」「られる」）は、朝日新聞の社説に比べて、工業新聞の記事では約 1.5 倍使われている。打消（「ない」「ぬ」）については、工業新聞の記事は朝日新聞の社説の約半分のみであった⁷⁾。ここで見る限り朝日新聞の社説と工業新聞の記事の文体を比較した場合、工業新聞の記事は、「受身」を多用する一方で「打消」を避ける傾向が観察された。

ただし、「断定」「推量」「伝聞」「様相」に関しては、朝日新聞の社説と工業新聞の記事に顕著

な差異は見られなかった。これは、朝日新聞の社説も工業新聞の記事もどちらも新聞記事であるということから考えると、事実を読者に伝達するという新聞記事特有の特徴とも考えられる。今後は、この点を踏まえて他のジャンルの文章と比較調査し、科学技術日本語の助動詞の用法に見られる特徴をより明らかにする必要がある。

これまでのところから、日本語の分析では、科学技術系の文章の特徴としてカタカナ語が多いことがわかったが、これは、科学系の専門用語には外来語が多いことが関係していると考えられる。また、科学系の文章に受身の助動詞が多く使われていたことは、動作主が表面に出ない受身という形を用いることで、事実を客観的に記述しようとしていると考えられる。打消の助動詞が少なかったことは、事実を記述する場合に、否定を使う修辭疑問文などあいまいな表現を避け、断定表現を選んで使っているからではないかと推察される。

3 科学技術英語の分析

英語の分析では、文の中核をなす動詞と助動詞を中心に調べ、科学技術系の英語の特徴を明らかにする。分析の対象データは、LOB コーパスと呼ばれる言語データを使用した。LOB コーパスとは、1978年に公開されたデータで、コンピュータによるイギリス英語の体系的な一次資料データベースである。LOB コーパスのタグ付きのものには、すべての単語に品詞情報が付けられている。

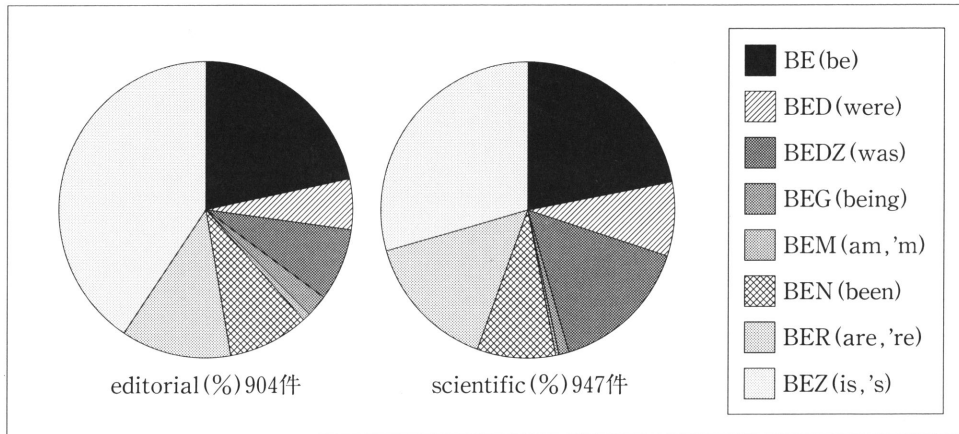
本研究では、LOB コーパスのタグ付きの Group B (社説) と Group J (科学系) を選んで分析した。それぞれのデータの大きさは230KBである。

3.1 BE 動詞と HAVE 助動詞

まず、Group B (社説) と Group J (科学系) のデータのそれぞれで、BE 動詞と HAVE 助動詞を計量分析した⁸⁾。表 5 は、社説と科学系のデータの中から BE 動詞をすべて検索し、その数を合計し割合を出した結果である。

〈表 5〉

BE 動詞のタグの種類	社説(%) 904件	科学系(%) 947件
1 BE (be)	21.5	22.2
2 BED (were)	5.5	7.7
3 BEDZ (was)	7.9	15.3
4 BEG (being)	2.7	1.6
5 BEM (am, 'm)	0.7	0.2
6 BEM (been)	9.0	8.2
7 BER (are, 're)	11.8	15.0
8 BEZ (is, 's)	41.0	29.8

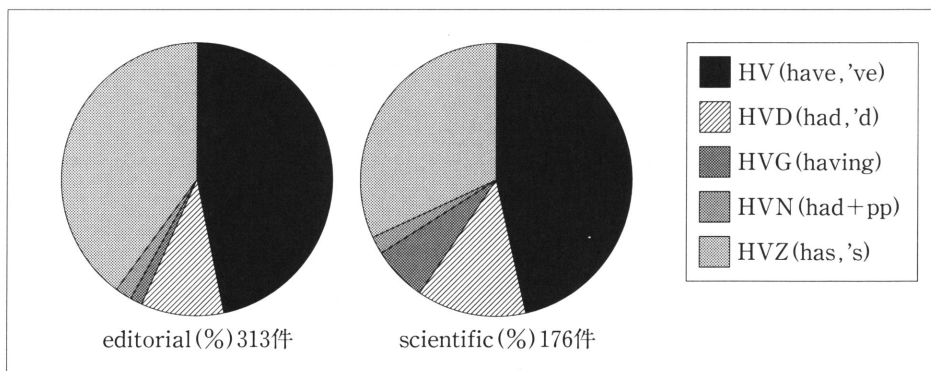


この結果から、社説でも科学系でも BE 動詞の中の BEZ (is, 's) がもっとも多く使われていることがわかる。特に社説の場合は 4 割以上も使われている。次に多いのは、社説、科学系のいずれの場合も BE (be) であった⁹⁾。

次に BE 動詞と同様の検索を HAVE 助動詞についても行った。社説と科学系のそれぞれのデータで HAVE 助動詞の HV (have, 've), HVD (had, 'd), HVG (having), HVN (had + pp), HVZ (has, 's) について検索し、数を合計し、割合を出した結果が表 6 である。

〈表 6〉

HAVE 助動詞のタグの種類	社説(%) 313件	科学系(%) 176件
1 HV (have, 've)	46.3	46.6
2 HVD (had, 'd)	10.2	12.5
3 HVG (having)	1.3	6.8
4 HVN (had + pp)	1.6	1.7
5 HVZ (has, 's)	40.6	32.4



結果は、社説も科学系も最も多いものがHV (have, 've) で、次がHVZ (has, 's) であった。HV (have, 've) とHVZ (has, 's) を合わせると社説、科学系いずれも9割近くになる。つまり、HAVE 助動詞のほとんどが単純現在完了を表すために使われている¹⁰⁾。

表5と表6の結果から見ると、社説と科学系のHAVE助動詞の種類のそれぞれの割合ではほとんど違いがないように見える。しかし、絶対数では違いが見られた。BE動詞の場合には、総計で社説は904件、科学系は947件と、ほとんど差違がない。これに対してHAVE助動詞の場合は、総計で科学系が176件に対して、社説では313件であった。つまりHAVE助動詞の使用が、科学系は社説の約半分しかなかったことになる¹¹⁾。

3.2 受動態の分析

次に受動態の分析を行った。一般に科学技術系英語の文章には受動態が多く使われるといわれている。これを計量的に実証できるか調査を試みた。LOBコーパスのGroup B (社説) と Group J (科学系) を対象に、受動態つまり「BE動詞+動詞の過去分詞形」を検索した。社説と科学系のそれぞれを検索した結果を合計したものを表7に示す¹²⁾。

〈表7〉

受動態 (BE 動詞+過去分詞形)	社説 (件)	科学系 (件)
	143	288

表7から社説の文章に比べて、科学技術系の文章では、約2倍の受動態が使われていることがわかる。

今回LOBコーパスの社説と科学系との比較分析でわかったことをまとめると以下の2点になる。

- 1) 科学系は社説と比べて、HAVE助動詞が約半分しか使われていなかった。また、BE動詞では両者に違いは見られなかった。
- 2) 科学系は社説と比べて、受動態が約2倍使われていた。

HAVE助動詞というのは、アスペクトの領域を表すものである。科学系の文章のように事実を記述する性質の文章では、アスペクト表現であるHAVE助動詞の使用は少ないと考えられる。また、科学系に受動態が多いのは、科学系の文章が客観的に表現しようとするため、動作主を明示することを避け、物事を主語にした記述になる傾向があるためと考えられる。

4. まとめ

本稿では、目的別語学教育のための基礎的研究として、科学系の文章と一般の文章との比較を行った。日本語の分析は、語彙の種類と動詞および助動詞の種類について計量的な分析をした。

その結果，科学系の文章にはカタカナが多用されること，助動詞の分析では，科学系に受身（「れる」「られる」）が1.5倍多くみられるが，打消（「ない」「ぬ」）は，社説に比べて約半分と少ない傾向にあること，動詞については差がなかったことが観察された。また，英語については，動詞と助動詞に焦点を当てて分析した。科学系は社説と比べて，HAVE 助動詞が約半分しか使われていないこと，また，科学系は社説と比べて，受動態が約2倍多く使われていることが観察された。その結果は以下のようにまとめられる。

日本語の比較（社説との比較）：科学系

カタカナの多用（約4.7倍）
受身の助動詞が約1.5倍
打消の助動詞が約半分

英語の比較（社説との比較）：科学系

受動態が約2倍
HAVE 助動詞が約半分

この様に，今回の調査で日本語，英語の両者に共通して，科学系の文章には受身の形が多く使われるということが確認された。冒頭でも述べたように，今回対象とした日本語の資料は，量・質の面から十分なものではなかった。今回の分析結果を踏まえて，今後は量的・質的に整った資料を対象に分析しなければならない。そして，今回の分析で指摘した日本語と英語の違いについてさらに詳しく研究していきたい。特に，英語の HAVE 助動詞で観察された完了の表現にあたる日本語の表現（「た」「ている」など）の調査や，日本語の打消の助動詞に対応する英語の否定表現（「not」「no」「never」など）の使われ方の調査によって，日本語，英語それぞれの科学系の文章の特徴を明らかにしていきたい。

注

- 1) 朝日新聞の社説は，1991年1月1日から10日まで，工業新聞の記事は，1994年1月20日から29日までの10日間のデータである。
- 2) 分析には，INDEXER の他に，VZ エディター（テキスト編集），LOTUS1-2-3（表計算）を使用した。INDEXER で形態素解析した出力結果を VZ エディターで読み込み，改行記号などの不要な記号などを取り除き，LOTUS1-2-3でソートした。
- 3) NTT 情報通信処理研究所の本木晴夫氏の協力を得て，INDEXER を中部大学の小森研究室のコンピュータ（NEC PC-9801RX）にインストールし，同研究所のご好意により INDEXER の使用を許可していただいている。この場をかりて感謝の意を表したい。
- 4) INDEXER の特徴は，入力した文章を解析し，形態素ごとに区切って，解析した結果を〈例1〉に示すように，出力するものであるが，もともとキーワードの作成のために作られていることから，固有名詞や熟語などを小さく区切ってしまう傾向があり，この点は今回のような語彙の分析をするためには，誤差となり不都合な点となった。

- 5) 1991年1月1日の社説の原文の一部は、以下のとおり。

国際政治と日本の体質強化【'91.1.1 朝刊 5頁 (全3278字)】

「宇宙から見る地球には、どこにも線は引っ張っていない。ほくたちの子ども、孫の時代には国境はなくなっているのかなと、しみじみ思います」——昨年暮れ日本人として初めて宇宙飛行をしたTBSの秋山特派員は、帰還後こう語った。いまに本当にそうになってほしいし、また現に、欧州共同体(EC)の市場統合のように、「国益至上主義」からの脱皮の兆しも見られぬではない。

- 6) 朝日新聞の社説と工業新聞の記事の動詞のデータの両者の相関係数は、 $r \approx 0.87$ で強い相関があった。
 7) 朝日新聞の社説と工業新聞の記事では、助動詞の使用に関して、相関係数は、 $r \approx 0.17$ で、ほとんど相関はなかった。
 8) たとえば、BE動詞を検索するCGREPのコマンドは、

```
>cgrep -l "BE¥>" edit.tag > editde
```

である。これは、edit.tagという名前のファイルの中にBEという記号(「¥>」は、単語の終わりを示す正規表現)があったらそれを含むすべての行を前後一行ずつの文脈も含めて抜き出して、editbeという名前のファイルに出力するという命令である。

- 9) 表5のデータの社説と科学系の両者の相関係数は、 $r \approx 0.81$ で、両者の間には強い相関がみられた。BE動詞に関して、社説と科学系の両者に割合としてほとんど差違は見られない。
 10) 表6の社説と科学系の両者の相関係数は、 $r \approx 0.78$ で強い相関がある。HAVE助動詞に関しても、社説と科学系の両者に割合としてほとんど差違は見られない。
 11) なぜ、HAVE助動詞が科学系に少なかったのか、詳しく分析しなければならない。そのためには検索結果の用例を個々に詳しく調査検討する必要がある。この分析については、次回の研究にゆだねることとする。
 12) CGREPで「BE動詞+動詞の過去分詞形」を抽出するコマンドは、

```
>cgrep -l "BE[ DGMNRZ][ Z].* VBN" edit.tag > editpp
```

である。これは、edit.tagという名前のファイルの中にBE動詞がタグのすべての種類BE, BED, BNDZ, BEG, BEM, BEN, BER, BEZとそれに後続する動詞の過去分詞形(VBN)があったらこれを含む行を前後一行ずつの文脈とともにeditppという名前のファイルに出力するという命令である。

参考文献

- Kimoto, H.(1991). Natural Language Processing and its Application to Japanese Database- the INDEXER System. Proceedings of the Third International Conference on Japanese Information in Science, Technology and Commerce (pp.447- 460).
 Komori, S. and H. Kimoto(1992). Tagged corpus and its application to teaching Japanese. Proceedings of the Second International Conference on Foreign Language Education and Technology (pp. 469- 474). LLA and IALL.
 Johansson, S. and K. Hofland(1989). Frequency Analysis of English Vocabulary and Grammar, 2 vols, Oxford: Clarendon.

資料

(1) Group B (社説) 中にある HAVE 助動詞の HVZ (has, 's) のデータの一例

以下に示すのは，CGREP で検索した語を含む行と前後の一行ずつを取り出して，SED でタグを取り除いたものである。行頭の数字は，Group B (社説) の行番号である。

BO5 57 and many a good trade unionist since then
BO5 57 has,, in the hope of
BO5 58 improving the lot of his fellow workers,,

BO5 62 - ^ now a new threat to those who toil
BO5 62 and spin has been developed by
BO5 63 a firm specialising in electronics in Los

BO5 198 bursting Japan..
BO5 199 ^ only one country in the world has more
BO5 199 people per square yard

BO6 36 bases all over western Europe..
BO6 37 ^ in the past year West Germany has
BO6 37 secured agreements for

(2) INDEXER の品詞コード表の一部 (助動詞の部分)

本文中の例 1 の 4 桁の数字は、品詞コードが以下のような分類で示される。

Ⅹ 1	意味	Ⅹ 2	意味	Ⅹ 3	意味
7	付属語	1	助	受身	1 れる
					2 られる
				使役	3 せる
					4 させる
					5 しめる
			動	希望	6 たい
					7 たがる
				打消	8 ない
					9 ぬ
		2	詞	過去	1 た
					2 だ
				断定	3 ます
					4 です
					5 だ
			推量		6 う
					7 よう
					8 まい
					9 べく、べき
		3	伝聞	1 ぞうだ	
				2 そうです	
			様相	3 らしい	
				4 ようだ	
				5 ようです	
				6 そうだ	
				7 そうです	
				8 みたいだ	